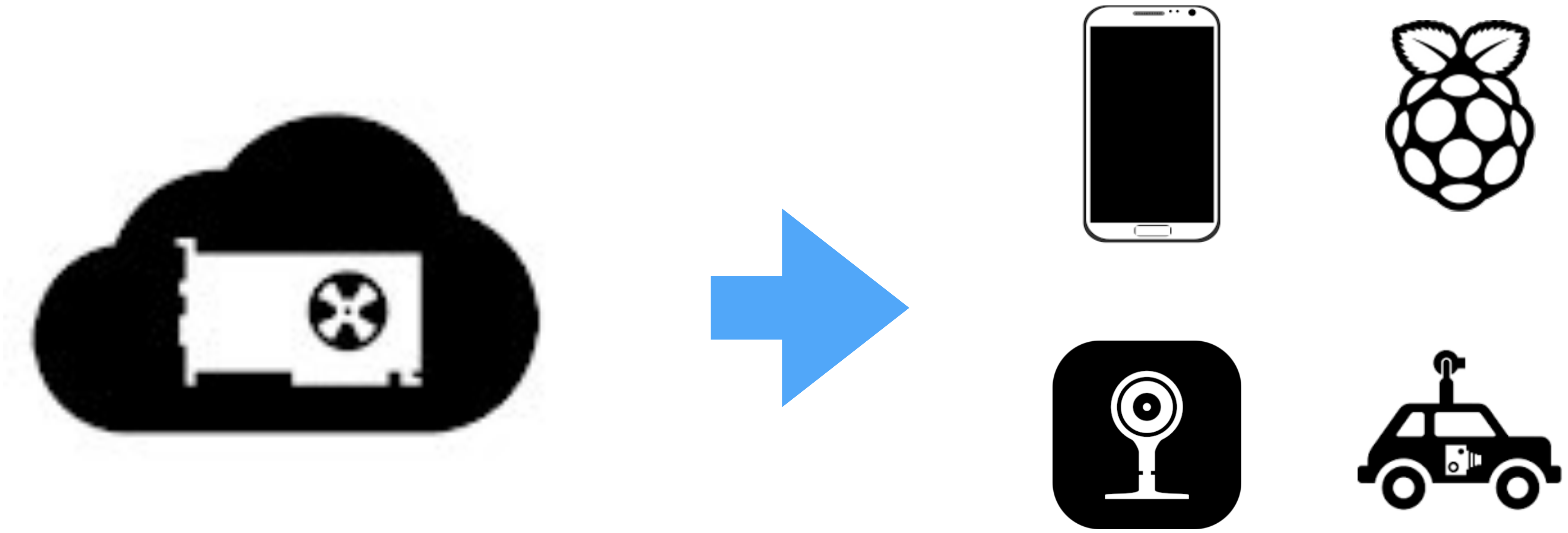
# Deep Learning from a Mobile* Perspective

Yangqing Jia
Research Scientist, Facebook

me@daggerfs.com

* mobile, on-device, embedded, IoT, anything without the cloud.

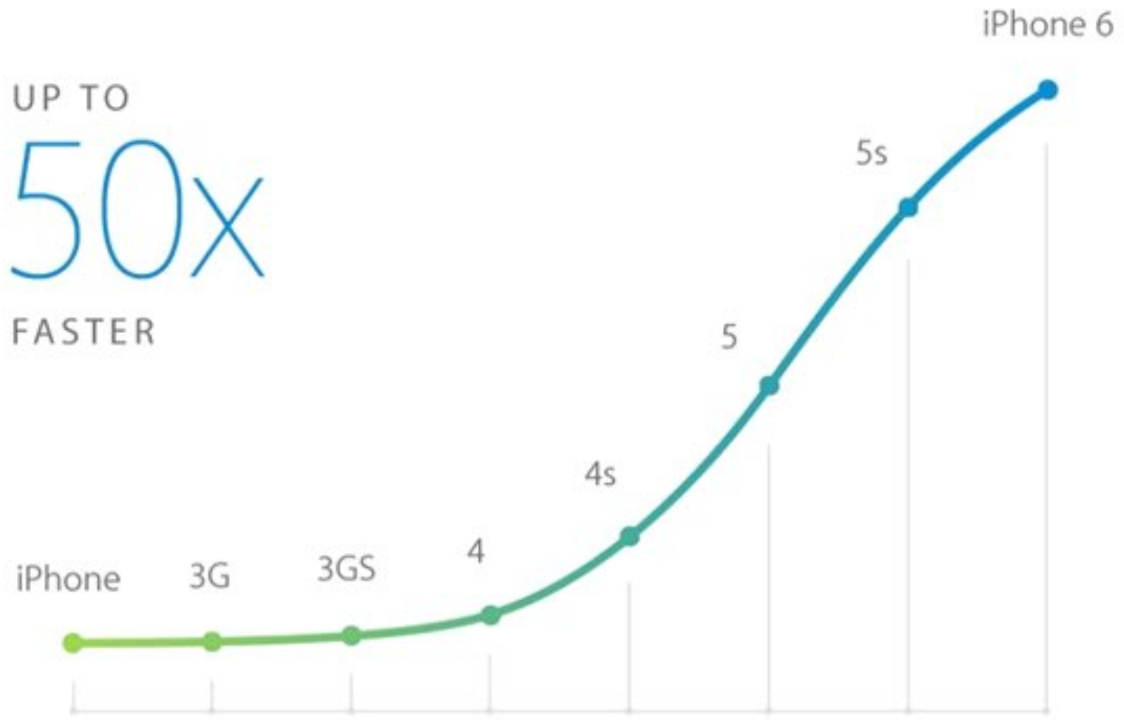
# Why Embedded Systems?



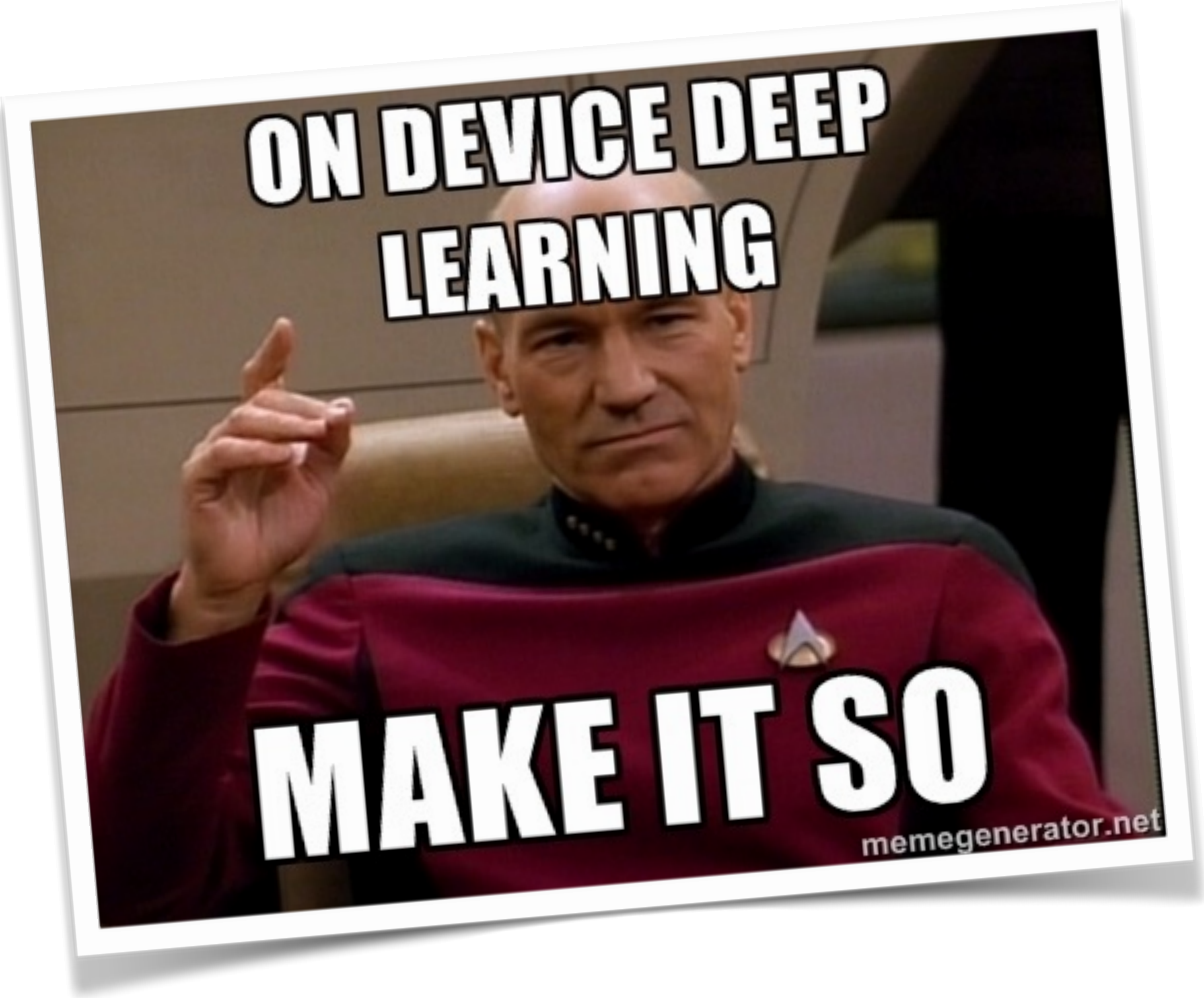
Expanding the capability of deep learning
to a wide range of applications

# Warp Speed

- Mobile is getting good!
  - Processor power
  - System architecture
  - Programming easiness
  - More applications



* Apple's processor growth over the years



But wait, captain...

# Challenges in an Embedded World

Speed

Power

Size

# Challenges in an Embedded World

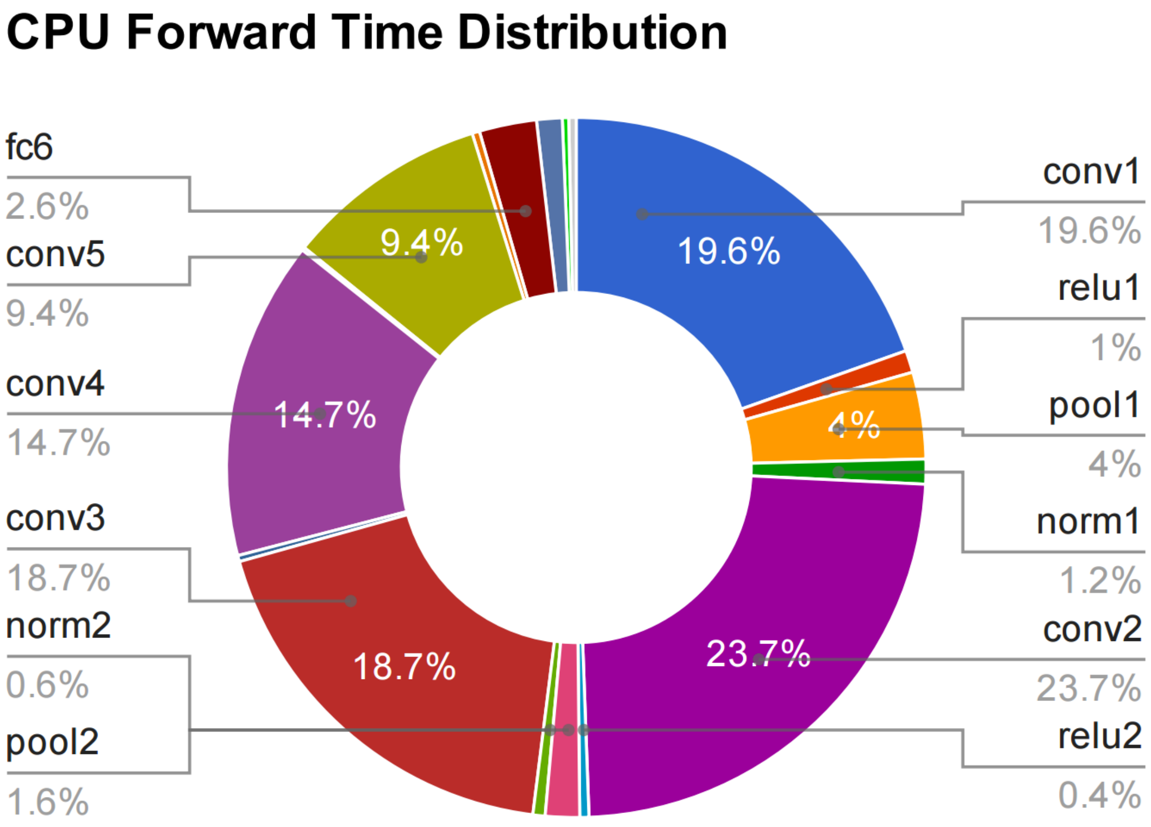
**Speed**

Power

Size

# Speed

- Why Gemm/Conv is at the heart of Deep Learning

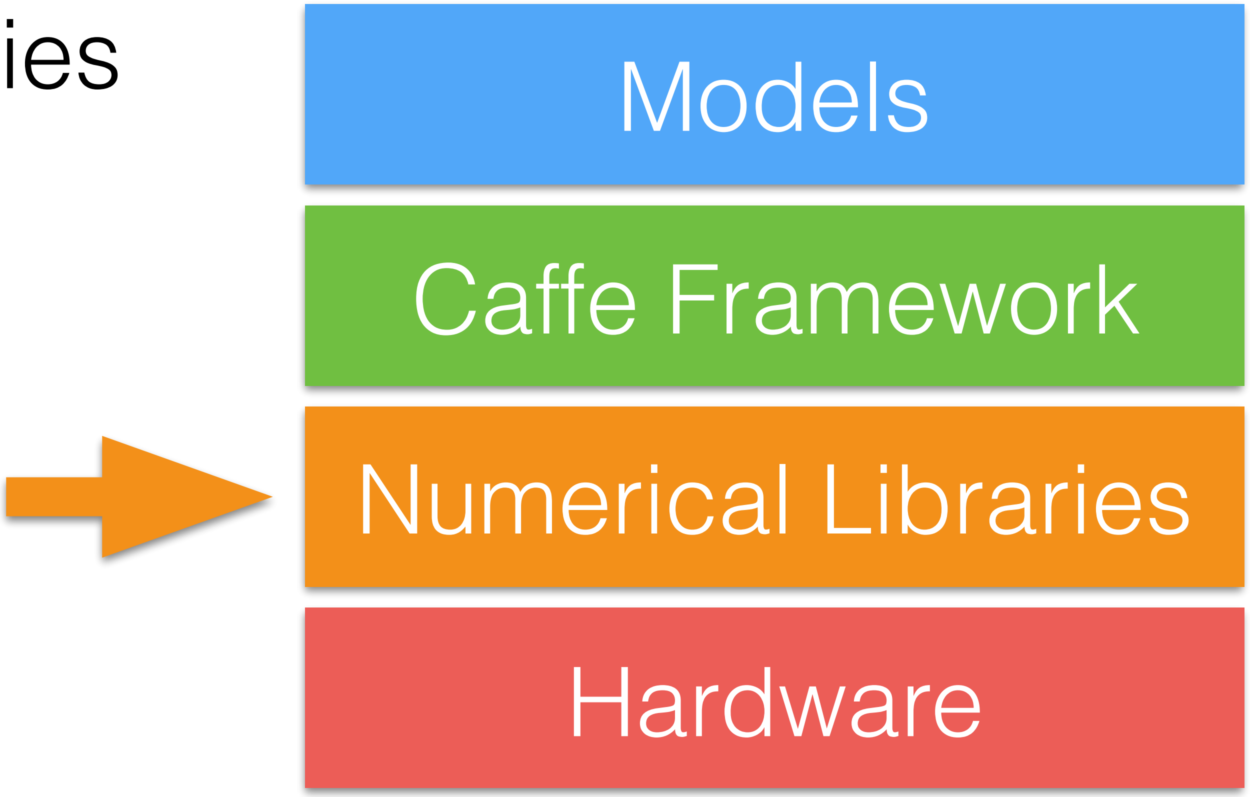


Source: UC Berkeley Thesis, Jia 2014

- Also, read Pete Warden's blog for more details

# Towards a Ludicrous Speed

- We need better libraries for numerical optimization!
  CuDNN
  MKL
  Eigen
  ....



# Challenges in an Embedded World

Speed

**Power**

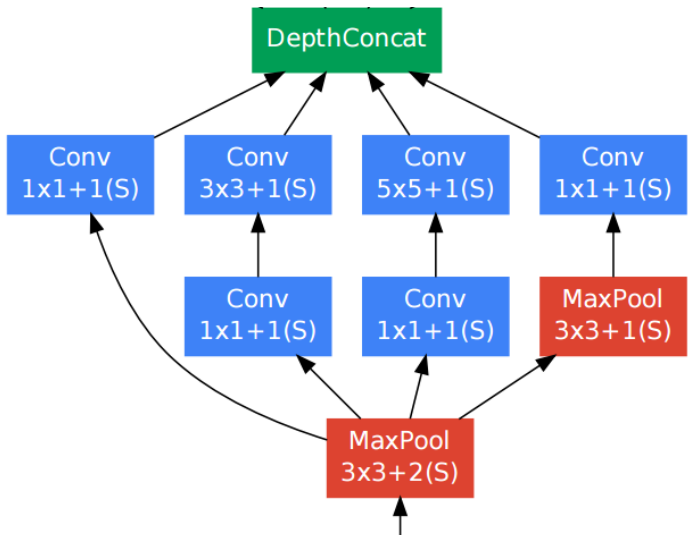
Size

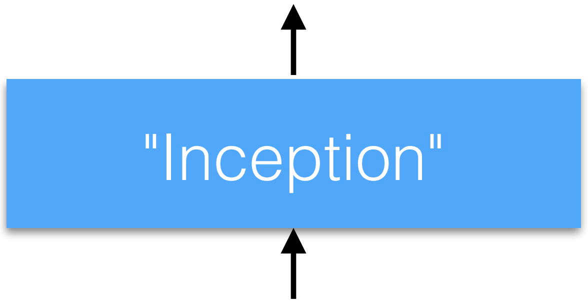
# DNN Demands, Battery Suffers

- DNNs are expensive
  Inception: 3 billion flops.
- Batteries don't last that long
  ~4 hrs (optimistic est.)
- Does not play well with thermo
  Burst in computation and intense use of cores

# The Dilemma between Modularity and Efficiency

- It's all about balance
- SW: Fine-grained vs Coarse operators
  RTC / JIT?
- HW: Efficiency vs Programmability
  FPGA / ASIC?




"Inception"

# Challenges in an Embedded World

Speed

Power

**Size**

# Putting Elephant in the Fridge

- DNN models are often very large

- We are getting better
  AlexNet: 240MB; Inception: 6MB
- But things are still wildly big for embedded
  Storage, bandwidth and memory limits

# There are (Potential) Ways

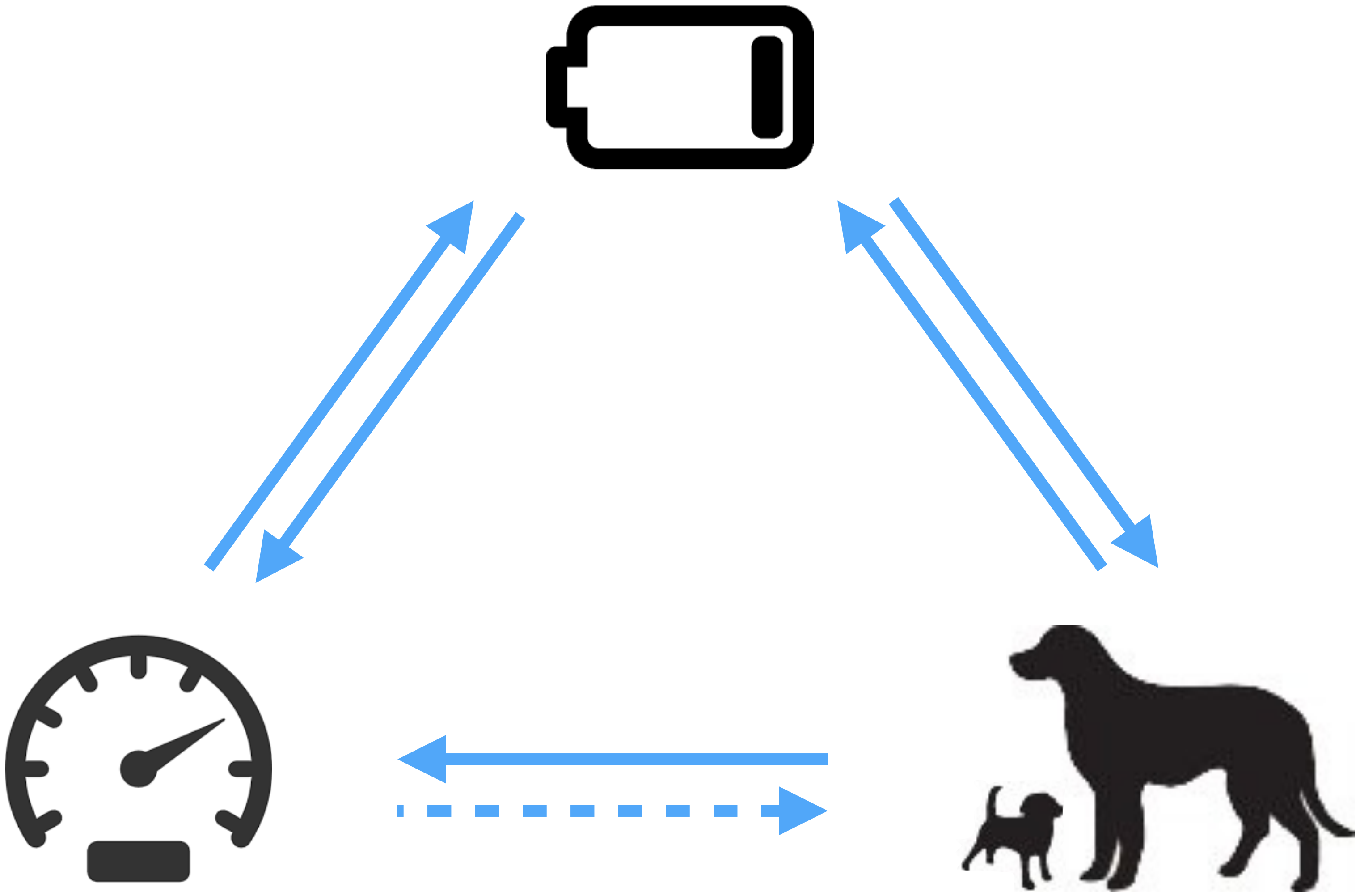
- Model Compression
  Compress the model but keep (approx.) its math

- Better Model Designs
  Inception, separable convolutions, etc.

- Quantization
  Float -> float16, int8, custom format...

- Distillation [Hinton 2015]
  Train small models to mimic big ones

# From a Joint View...



“To boldly go where no one has gone before.”